Inferring Meta-Covariates in Classification

Keith Harris, Lisa McMillan and Mark Girolami

Inference Group, Department of Computing Science, University of Glasgow, UK
{keithh,lisa,girolami}@dcs.gla.ac.uk
http://www.dcs.gla.ac.uk/inference

Abstract. This paper develops an alternative method for gene selection that combines model based clustering and binary classification. By averaging the covariates within the clusters obtained from model based clustering, we define "meta-covariates" and use them to build a probit regression model, thereby selecting clusters of similarly behaving genes, aiding interpretation. This simultaneous learning task is accomplished by an EM algorithm that optimises a single likelihood function which rewards good performance at both classification and clustering. We explore the performance of our methodology on a well known leukaemia dataset and use the Gene Ontology to interpret our results.

Key words: Gene selection, clustering, classification, EM algorithm, Gene Ontology.

1 Introduction

In this paper, we develop a procedure for potentially improving the classification of gene expression profiles through coupling with the method of model based clustering. Such DNA microarray data typically consists of several thousands of genes (covariates) and a much smaller number of samples. Analysing this data is statistically challenging, as the covariates are highly correlated, which results in unstable parameter estimates and inaccurate prediction. To alleviate this problem, we use the averages of covariate clusters, rather than all the original covariates, to classify DNA samples. The advantage of this approach over using a sparse classification model [1, 2] is that we can extract a much larger subset of genes with essential predictive power and partition this subset into groups, within which the genes are similar.

An overview of our procedure that combines model based clustering and binary classification is as follows. By averaging the features within the clusters obtained from a Gaussian mixture model [3, 4], we define "superfeatures" or "meta-covariates" and use them in a probit regression model, thereby attaining concise interpretation and accuracy. Similar ideas, from a non-Bayesian two-step perspective, have been looked at by Hanczar et al. [5] and Park et al. [6]. With our simultaneous procedure, the clusters are formed considering the correlation of the predictors with the response in addition to the correlations among the predictors. The proposed methodology should have wide applicability in areas such as gene selection and proteomic biomarker selection.

The rest of this paper is organized as follows: in Sect. 2 we introduce our meta-covariate classication model and provide an EM algorithm for learning the parameters of our model from data. In Sect. 3 we illustrate our method with a DNA microarray data example and use the Gene Ontology (GO) to interpret our results. Section 4 discusses the conclusions we draw from our experimental results. Finally, Appendix A gives the full details of our model and shows the derivation of our EM algorithm.

2 Methodology

2.1 Model

In the following discussion, we will denote the $N \times D$ design matrix as $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N]^T$ and the $N \times 1$ vector of associated response values as $\boldsymbol{t}$ where each element $t_n \in \{-1, 1\}$. The $K \times N$ matrix of clustering mean parameters θ_{kn} is denoted by θ. We represent the $K \times 1$-dimensional columns of θ by $\boldsymbol{\theta}_n$ and the corresponding $N \times 1$-dimensional rows of θ by $\boldsymbol{\theta}_k$. The $D \times K$ matrix of clustering latent variables z_{dk} is represented as Z. The $K \times 1$ vector of regression coefficients is denoted by $\boldsymbol{w}$. Finally, we denote the $N \times 1$ vector of classification auxiliary variables by $\boldsymbol{y}$.

The graphical representation of the conditional dependency structure in the meta-covariate classification model is shown in Fig. 1. From Fig. 1 we see that the joint distribution of the meta-covariate classification model is given by

$$p(\boldsymbol{t}, \boldsymbol{y}, X, \theta, \boldsymbol{w}) = p(\boldsymbol{t}, \boldsymbol{y}|\theta, \boldsymbol{w})p(X|\theta)p(\theta)p(\boldsymbol{w}). \tag{1}$$

The distribution $p(X|\theta)$ is the likelihood contribution from our clustering model, which we chose to be a normal mixture model with equal weights and identity covariance matrices. Similarly, $p(\boldsymbol{t}, \boldsymbol{y}|\theta, \boldsymbol{w})$ is the likelihood contribution from our classification model, which we chose to be a probit regression model whose covariates are the means of each cluster, that is, $\boldsymbol{\theta}_k$, $k = 1, \ldots, K$. Finally, the model was completed by specifying vague normal priors for θ and $\boldsymbol{w}$. Full details of our model along with the derivation of the following EM algorithm that we used for inference is given in Appendix A.

2.2 Summary of the EM Algorithm

Given the number of clusters K, the goal is to maximise the joint distribution with respect to the parameters (comprising the means of the clusters and the regression coefficients).

1. Initialise θ, $\boldsymbol{w}$, the responsibilities $\gamma(z_{dk})$ and $E(\boldsymbol{y})$, and evaluate the initial value of the log likelihood.

2. E-step. Evaluate:

$$\gamma(z_{dk}) = \frac{\exp\left\{-\frac{1}{2}\|\boldsymbol{x}_d - \boldsymbol{\theta}_k\|^2\right\}}{\sum_{j=1}^{K} \exp\left\{-\frac{1}{2}\|\boldsymbol{x}_d - \boldsymbol{\theta}_j\|^2\right\}} \tag{2}$$

and

$$E(y_n) = \begin{cases} \boldsymbol{w}^T\boldsymbol{\theta}_n + \frac{\phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)}{1-\Phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)} & \text{if } t_n = 1 \\ \boldsymbol{w}^T\boldsymbol{\theta}_n - \frac{\phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)}{\Phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)} & \text{otherwise.} \end{cases} \tag{3}$$

3. M-step. Evaluate:

$$\boldsymbol{\theta}_k = \frac{\left(E(\boldsymbol{y}) - \theta^T\boldsymbol{w}_{-k}\right) w_k + X\boldsymbol{\gamma}_k + \frac{1}{h}\boldsymbol{\theta}_0}{w_k^2 + \sum_{d=1}^{D}\gamma(z_{dk}) + \frac{1}{h}} \tag{4}$$

and

$$\boldsymbol{w} = \left(\theta\theta^T + \frac{1}{l}I\right)^{-1}\theta E(\boldsymbol{y}). \tag{5}$$

After updating $\boldsymbol{w}$ in this manner, set the first component of the vector to 1, so that the model is identifiable.
4. Evaluate the log likelihood and check for convergence. If the convergence criterion is not satisfied return to step 2.

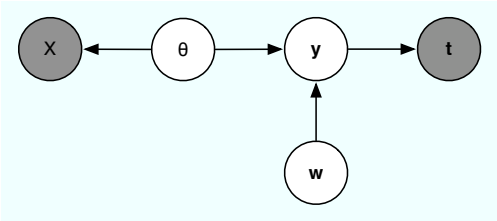

Fig. 1. Graphical representation of the conditional dependencies within the meta-covariate classification model.

3 Experimental Results - Acute Leukemia Data

3.1 Data Description

A typical application where clustering and classification have become common tasks is the analysis of DNA microarray data, where thousands of gene expression levels are monitored on a few samples of different types. We thus decided to illustrate our proposed methodology for inferring meta-covariates in classification with the widely analysed leukaemia microarray dataset of Golub et al. [7], which was downloaded from the Broad Institute Website[1]. Bone marrow or peripheral blood samples were taken from 72 patients with either acute myeloid leukaemia (AML) or acute lymphoblastic leukaemia (ALL). Gene expression levels were

[1] `http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=43`

measured using Affymetrix high-density oligonucleotide arrays containing 7129 probes for 6817 human genes. Following the experimental setup of the original paper, the dataset was split into a training set of 38 samples of which 27 are ALL and 11 are AML, and a test set of 34 samples, 20 ALL and 14 AML. The data was preprocessed as recommended in [8]: (a) thresholding, floor of 100 and ceiling of 16000; (b) filtering, exclusion of probes with max/min ≤ 5 and (max $-$ min) $\leq$ 500; (c) base 10 logarithmic transformation; and (d) standardising, so that each sample has mean 0 and variance 1. This left us with 3571 probes for analysis. Finally, GO annotations for the appropriate gene chip (Hu6800) were obtained via the Affymetrix NetAffx analysis centre[2].

3.2 Results and Discussion

EM algorithm results. Figure 2 shows the minimum and mean test error from 200 runs of our EM algorithm for different values of the number of clusters K. It should be noted that we used the K-means clustering algorithm to initialise the matrix of clustering mean parameters θ, while the other parameters were initialised randomly. We see from Fig. 2 that on average the algorithm performs best for around 15 to 25 clusters, with the best case yielding an average test error rate of 9.93% for $K = 21$ clusters. We also see that for $K = 21$ clusters, the run that achieved the highest likelihood value also achieved the minimum test error of 2.94%, that is, just one misclassification in the test set. The predictions from the highest likelihood model with $K = 21$ clusters thus appear to improve predictions made by Golub et al. [7], who made five misclassifications on the test set, and is competitive with the methods of Lee et al. [1] and Bae and Mallick [2], who misclassified one and two test samples, respectively. We will now use the Gene Ontology to interpret the results from this model.

GO analysis. Table 1 describes each of the 21 probe clusters, with respect to the number of probes allocated to the cluster; the number of *control* probes allocated to the cluster; its regression coefficient (w); its rank by descending absolute regression coefficient; and the number of genes represented by the probe set. The number of unique Entrez Gene IDs (as obtained from NetAffx) was used to count the number of unique genes.

22 of the 59 controls on the microarray survive the initial filtering process (all 22 of these are endogenous controls). Control probes, by design, should not be functionally significant. It is therefore encouraging that most (63.64%) of the control probes belong to the four least influential clusters (with respect to abs(w)): clusters 9 ($w = -0.15$, ranked 21st), 8 ($w = 0.16$, ranked 20th), 2 ($w = 0.22$, ranked 19th) and 7 ($w = -0.37$, ranked 18th). Furthermore, cluster 8 – the cluster with the lowest absolute regression coefficient – contains only four probes, all of which are control probes. It should be noted that six control probes do occur in the ten 'significant' clusters; the extent to which these probes are appropriate controls should be investigated further.

[2] `http://www.affymetrix.com/analysis/index.affx`

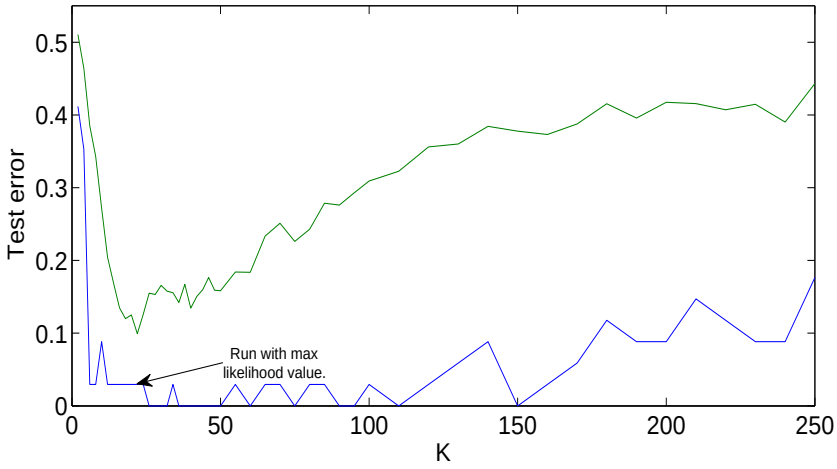

Fig. 2. Minimum and mean test error after 200 runs of the EM algorithm.

Table 1. The best clusters ($K = 21$).

Cluster	Probes	Controls	w	rank(abs(w))	Genes
1	20	0	1.00	10	16
2	486	4	0.22	19	412
3	20	0	-1.22	8	20
4	253	0	-1.88	7	230
5	182	0	0.55	15	173
6	240	1	-3.08	4	199
7	110	2	-0.37	18	99
8	60	4	0.16	20	50
9	4	4	-0.15	21	1
10	230	0	-2.66	5	214
11	189	1	-1.10	9	166
12	210	1	0.88	12	183
13	228	0	0.79	13	200
14	230	0	0.55	16	187
15	61	0	3.87	1	56
16	240	0	3.21	3	204
17	213	0	-0.50	17	205
18	17	0	-0.95	11	16
19	267	1	0.75	14	235
20	101	1	-3.79	2	85
21	210	3	2.46	6	175

The clusters are reasonably well balanced, with most clusters containing approximately 200 genes. The largest and smallest clusters (numbers 2 and 9 respectively) have small regression coefficients, indicating that they have limited influence on the classifier.

Using $w = 1$ as a baseline, ten clusters (numbers 15, 20, 16, 6, 10, 21, 4, 3, 11, 1) are sufficiently weighted to be of interest (these ten clusters will be described as the 'significant' clusters). The aim of this work is to assess whether there is any biological significance in the clustering of the probes (or genes): the expectation is that genes clustered together will be carrying out a similar function or functions. As such, GO annotations from the molecular function aspect of the GO were used.

The total number of occurrences for each GO term across all genes in a cluster was calculated. By comparing this to the occurrences for each GO term across the entire chip and using the hypergeometric distribution, we can calculate the probability that the terms were encountered by chance. By comparing the occurrence of the GO term in the cluster and the entire chip, we can describe it as over- or under-represented in the gene cluster.

Cluster 15, $w = 3.87$. Most noticeably, metal ion (and specifically zinc ion) annotations are under-represented in this gene cluster. Further, nucleotide and nucleic acid binding are seen less often than would be expected. Several very specific terms are found enriched in this gene cluster; of particular interest is a cluster of three enzyme inhibitor activity subterms.

Cluster 20, $w = -3.79$. There is a concentration of very specific transmembrane transporter activities and oxidoreductase terms. Unlike the previous cluster, protein kinase activity is under-represented; nucleic acid binding is over-represented and receptor activity is under-represented in this cluster.

Cluster 16, $w = 3.21$. In this cluster, zinc ion binding is over-represented, unlike in clusters 15 and 20 (where the term was under-represented and not significant respectively). Also interesting is the overrepresentation of the "damaged DNA binding" term - particularly relevant in the context of cancer. Like cluster 15, several general receptor binding terms are over-represented. A small cluster of pyrophosphatase subterms are also over-represented.

Cluster 6, $w = -3.08$. Several metal ion binding terms are over-represented here, including calcium and zinc, and most interestingly – particularly in the context of leukaemia, cancer of the blood – heme binding. Again, several receptor binding and activity terms are over-represented.

Cluster 10, $w = -2.66$. Most noticeably, a small cluster of under-represented terms describe signal transducer activity and several kinds of receptor activities. This is an area of the Gene Ontology that was enriched in clusters 15, 16 and 6 and under-represented in cluster 20. There is significant enrichment of DNA binding terms (specifically DNA topoisomerases).

Cluster 21, $w = 2.46$. Cluster 21 has the most extensive coverage and deepest annotation of the ten significant clusters, despite being of comparable size to many others (e.g., 16, 6, 10, 4 and 11). In addition, none of the significant annotations are seen less than would be expected: they are all enriched in this cluster. Multiple metal ion binding terms are enriched here as are DNA binding, receptor activity and kinase activity.

Cluster 4, $w = -1.88$. Cluster 4 is enriched for several transcription regulation terms, kinase activities, and DNA and nucleotide binding. Here, enzyme regulator activities are under-represented.

Cluster 3, $w = -1.22$. The genes in cluster 3 are enriched for receptor activity and a specific receptor activity: fibroblast growth factor receptor activity. Again, receptor binding and activity terms are over-represented and metal ion terms are under-represented. There is enrichment of a specific enzyme activator activity, apoptotic protease activator activity, of particular interest in the context of cancer.

Cluster 11, $w = -1.10$. A cluster of signal transducer activity/receptor activities are under-represented here; similar to patterns observed in clusters 20, 4 and 10. There are fewer metal (iron, calcium and zinc) ion binding terms and protein kinase annotations than would be expected by chance.

Cluster 1, $w = 1.00$. Cluster 1 defines the 'baseline' for regression model coefficients. This cluster is enriched for ion binding (including iron, ferrous, haem and haemoglobin), ferrochelatase and oxygen transporter activity, significant in the context of leukaemia.

Table 2 describes each of the ten significant clusters with respect to an annotation profile, which considers over-representation and under-representation of metal ion binding terms; DNA or RNA binding terms; receptor activity terms; enzyme regulation terms; receptor binding terms; kinase activity terms; transmembrane transport terms and transcriptional regulation terms.

It is clear that none of the clusters are identical with respect to this profile. Receptor activity terms and metal ion binding terms are more often over-representated in the gene clusters with positive regression coefficients, and more often under-represented in the gene clusters with negative regression coefficients.

Comparison to other methods. In their original paper, Golub et al. [7] identified 50 genes that were highly correlated with the AML/ALL class distinction. 68% of these genes are assigned to a cluster with an absolute regression coefficient of ≥ 1. Cluster 15, the top ranking cluster with respect to absolute regression coefficient, contains six of these genes and cluster 20, the next most influential cluster, contains four of these genes. Surprisingly, eight genes are found in cluster 5, which has a low regression coefficient ($w = 0.55$).

Table 2. Summary of cluster annotations.

Cluster	w	MIB	D/RB	RA	ER	RB	KA	TMT	TRR
15	3.87	n	n	y	y	y	y	y	
16	3.21	y	y	y	y	y		y	$\sim$
21	2.46	y	y	y		y	y	y	y
1	1.00	y		y					
11	-1.10	n		n				y	
3	-1.22	n		y	y	y	y	y	
4	-1.88		y	n	y		y		y
10	-2.66	n	y	n			n	y	
6	-3.08	y		y					y
20	-3.79		y	n		y	n	y	

MIB = metal ion binding; D/RB = DNA or RNA binding; RA = receptor activity; ER = enzyme regulation; RB = receptor binding; KA = kinase activity; TMT= transmembrane transport; TRR = transcription regulation. y indicates over-representation; n indicates under-representation; $\sim$ indicates conflicting results.

More recently, Lee et al. [1] identified 27 genes as informative, using a Bayesian method for variable selection. In this more refined set, eight (29.63%) of the genes belong to the most influential cluster (15). In a follow up study where sparsity was imposed on the priors, Bae and Mallick [2] identified 10 genes using various models. Here, three genes are found in cluster 15 and two genes are found in cluster 20, and only two genes are mapped to clusters with an absolute regression coefficient < 1.

Three genes are identified by all three methods [1, 2, 7]: Cystatin C, Zyxin and CF3 (transcription factor 3). CF3 is assigned to cluster 5, a comparatively weakly informative cluster; however, both Zyxin and Cystatin C are assigned to cluster 15, the most influential cluster in the regression model.

4 Conclusions

The method is successful in assigning limited influence to control probes. The clustering of probes reflects functional differences between the genes that they represent. Furthermore, enrichment of metal ion binding and receptor activity annotations appear to correspond with the sign of the regression coefficients; that is, clusters with positive regression coefficients are more often enriched for such annotations, while clusters with negative regression coefficients are often under-represented by such annotations.

In a comparison with methods of variable selection in the same dataset, genes important in the discrimination between AML and ALL tend to belong to clusters with high absolute regression coefficients in the model; this is particularly true as the variable selection methods become more sophisticated and fewer genes are found to be significant. Of the three genes that are common in three

different analyses of these data, two (Zyxin and Cystatin C) are assigned to the most influential cluster in our model.

Our experimental results thus indicate that our EM algorithm approach of inferring meta-covariates in classification is a promising new methodology with wide applicability. Moreover, the approach can be naturally extended to multi-class classification and to incorporate sparsity by employing an Inverse Gamma prior on the variance of the regression coefficients. Future research will focus on developing a Bayesian sampler for the "meta-covariate" classification model, possibly using reversible jump Markov chain Monte Carlo or an infinite mixture model to infer directly from the data the optimal number of clusters.

Acknowledgements. K. Harris & M. Girolami are supported by the Engineering and Physical Sciences Research Council (EPSRC) grant EP/F009429/1 - Advancing Machine Learning Methodology for New Classes of Prediction Problems. M. Girolami is funded by an EPSRC Advanced Research Fellowship EP/E052029/1. L. McMillan is funded by a grant from SHEFC SRDG.

References

1. Lee, K.E., Sha, N., Dougherty, E.R., Vannucci, M., Mallick, B.K.: Gene selection: a Bayesian variable selection approach. Bioinformatics **19**(1) (January 2003) 90–97
2. Bae, K., Mallick, B.K.: Gene selection using a two-level hierarchical Bayesian model. Bioinformatics **20**(18) (July 2004) 3423–3430
3. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer (2006)
4. Fraley, C., Raftery, A.E.: Bayesian regularization for normal mixture estimation and model-based clustering. Journal of Classification **24**(2) (September 2007) 155–181
5. Hanczar, B., Courtine, M., Benis, A., Henegar, C., Clément, K., Zucker, J.D.: Improving classification of microarray data using prototype-based feature selection. SIGKDD Explorations **5**(2) (December 2003) 23–30
6. Park, M.Y., Hastie, T., Tibshirani, R.: Averaged gene expressions for regression. Biostatistics **8**(2) (April 2007) 212–227
7. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439) (October 1999) 531–537
8. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association **97**(457) (March 2002) 77–87

A Derivation of the EM Algorithm

A.1 Regression

Modelling. In the following subsection, $\boldsymbol{y}$ denotes an $N \times 1$ continuous response vector.

Joint distribution.

$$p(\boldsymbol{y}, X, \theta, \boldsymbol{w}) = p(\boldsymbol{y}|\theta, \boldsymbol{w})p(X|\theta)p(\theta)p(\boldsymbol{w}). \tag{6}$$

Regression model.

$$y_n = \boldsymbol{w}^T\boldsymbol{\theta}_n + \epsilon_n \text{ where } \epsilon_n \sim \mathcal{N}(0,1). \tag{7}$$

$$\Rightarrow p(\boldsymbol{y}|\theta, \boldsymbol{w}) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\theta}_n, \boldsymbol{w}) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_n - \boldsymbol{w}^T\boldsymbol{\theta}_n)^2}. \tag{8}$$

$$\Rightarrow \log p(\boldsymbol{y}|\theta, \boldsymbol{w}) = -\frac{1}{2}\sum_{n=1}^{N}(y_n - \boldsymbol{w}^T\boldsymbol{\theta}_n)^2 - \frac{N}{2}\log(2\pi). \tag{9}$$

Clustering model. Normal mixture model with equal weights and identity covariance matrices.

$$\Rightarrow p(\boldsymbol{x}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{x}|\boldsymbol{\theta}_k, I). \tag{10}$$

From [3] we have that:

$$\log p(X|\theta) \geq \mathcal{L}(q,\theta) = \sum_{Z} q(Z)\log\left\{\frac{p(X,Z|\theta)}{q(Z)}\right\}, \tag{11}$$

where Z is a $D \times K$ matrix of latent variables with rows $\boldsymbol{z}_d^T$ such that $\boldsymbol{z}_d$ is a K-dimensional binary random variable having a 1 of K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0, and $q(Z)$ is a distribution defined over the latent variables.

$$\Rightarrow \log p(X|\theta) \geq \sum_{Z} p(Z|X,\theta^{\text{old}})\log p(X,Z|\theta) - \sum_{Z} p(Z|X,\theta^{\text{old}})\log p(Z|X,\theta^{\text{old}}) \tag{12}$$

$$= Q(\theta,\theta^{\text{old}}) + \text{const.} \tag{13}$$

$$p(X,Z|\theta) = \prod_{d=1}^{D}\prod_{k=1}^{K}\left(\frac{1}{K}\right)^{z_{dk}}\mathcal{N}(\boldsymbol{x}_d|\boldsymbol{\theta}_k, I)^{z_{dk}}. \tag{14}$$

$$\Rightarrow E_Z[\log p(X,Z|\theta)] \geq -\frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}E(z_{dk})\sum_{n=1}^{N}(x_{nd} - \theta_{kn})^2 + \text{const.} \tag{15}$$

Prior distributions.

$$p(\theta) = \prod_{k=1}^{K}\mathcal{N}(\boldsymbol{\theta}_k|\boldsymbol{\theta}_0, hI), \tag{16}$$

where each element of $\boldsymbol{\theta}_0$ is set to the corresponding covariate interval midpoint and h is chosen arbitrarily large in order to prevent the specification of priors that don't overlap with the likelihood and allow for mixtures with widely different component means.

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, lI). \tag{17}$$

E-step.

$$E(z_{dk}) = \gamma(z_{dk}) = \frac{\sum_{z_{dk}} z_{dk} \left[\frac{1}{K}\mathcal{N}(\boldsymbol{x}_d|\boldsymbol{\theta}_k, I)\right]^{z_{dk}}}{\sum_{z_{dj}} \left[\frac{1}{K}\mathcal{N}(\boldsymbol{x}_d|\boldsymbol{\theta}_j, I)\right]^{z_{dj}}} \tag{18}$$

$$= \frac{\exp\left\{-\frac{1}{2}\|\boldsymbol{x}_d - \boldsymbol{\theta}_k\|^2\right\}}{\sum_{j=1}^{K} \exp\left\{-\frac{1}{2}\|\boldsymbol{x}_d - \boldsymbol{\theta}_j\|^2\right\}}. \tag{19}$$

M-step.

$$\log p(\boldsymbol{y}, X, \theta, \boldsymbol{w}) \geq -\frac{1}{2}\sum_{n=1}^{N}\left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right)^2 - \frac{1}{2}\sum_{d=1}^{D}\sum_{k=1}^{K}\gamma(z_{dk})\sum_{n=1}^{N}(x_{nd}-\theta_{kn})^2$$
$$- \frac{1}{2h}\sum_{k=1}^{K}\sum_{n=1}^{N}(\theta_{kn} - \theta_{0n})^2 - \frac{1}{2l}\sum_{k=1}^{K} w_k^2 + \text{const.} \tag{20}$$

$$\frac{\partial \log p(\boldsymbol{y}, X, \theta, \boldsymbol{w})}{\partial \theta_{kn}} = \left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right) w_k + \sum_{d=1}^{D}\gamma(z_{dk})(x_{nd}-\theta_{kn}) - \frac{1}{h}(\theta_{kn}-\theta_{0n}) = 0. \tag{21}$$

$$\Rightarrow \boldsymbol{\theta}_k = \frac{\left(\boldsymbol{y} - \theta^T \boldsymbol{w}_{-k}\right) w_k + X\boldsymbol{\gamma}_k + \frac{1}{h}\boldsymbol{\theta}_0}{w_k^2 + \sum_{d=1}^{D}\gamma(z_{dk}) + \frac{1}{h}}, \tag{22}$$

where $\boldsymbol{w}_{-k}$ is $\boldsymbol{w}$ with the k^{th} element set to 0 and $\boldsymbol{\gamma}_k$ is the $D \times 1$-dimensional column of the $D \times K$ matrix of responsibilities $[\gamma(z_{dk})]$.

$$\frac{\partial \log p(\boldsymbol{y}, X, \theta, \boldsymbol{w})}{\partial w_k} = \sum_{n=1}^{N}\left(y_n - \sum_{k'=1}^{K} w_{k'}\theta_{k'n}\right)\theta_{kn} - \frac{1}{l}w_k = 0. \tag{23}$$

$$\Rightarrow \boldsymbol{w} = \left(\theta\theta^T + \frac{1}{l}I\right)^{-1}\theta\boldsymbol{y}. \tag{24}$$

A.2 Extension to Binary Classification

Modelling.

Joint distribution. The joint distribution now becomes

$$p(\boldsymbol{t}, \boldsymbol{y}, X, \theta, \boldsymbol{w}) = p(\boldsymbol{t}, \boldsymbol{y}|\theta, \boldsymbol{w})p(X|\theta)p(\theta)p(\boldsymbol{w}). \tag{25}$$

Classification model.

$$t_n = \begin{cases} 1 & \text{if } y_n > 0 \\ -1 & \text{otherwise.} \end{cases} \tag{26}$$

$$y_n = \boldsymbol{w}^T \boldsymbol{\theta}_n + \epsilon_n \text{ where } \epsilon_n \sim \mathcal{N}(0,1). \tag{27}$$

$$\Rightarrow p(\boldsymbol{t}, \boldsymbol{y}|\theta, \boldsymbol{w}) = \prod_{n=1}^{N} p(t_n, y_n|\boldsymbol{\theta}_n, \boldsymbol{w}) \tag{28}$$

$$= \prod_{n=1}^{N} p(t_n|y_n) p(y_n|\boldsymbol{\theta}_n, \boldsymbol{w}) \tag{29}$$

$$= \prod_{n=1}^{N} p(t_n|y_n) \mathcal{N}\big(y_n|\boldsymbol{w}^T\boldsymbol{\theta}_n, 1\big)\,, \tag{30}$$

where

$$p(t_n|y_n) = \begin{cases} \delta(y_n > 0) & \text{if } t_n = 1 \\ \delta(y_n \leq 0) & \text{otherwise.} \end{cases} \tag{31}$$

E-step. Then, by taking logarithms and applying Jensen's inequality, we obtain the following result:

$$E_{\boldsymbol{y}}[\log p(\boldsymbol{t}, \boldsymbol{y}|\theta, \boldsymbol{w})] \geq \sum_{n=1}^{N} \log \left[p(t_n|E(y_n)) \mathcal{N}\big(E(y_n)|\boldsymbol{w}^T\boldsymbol{\theta}_n, 1\big)\right]. \tag{32}$$

$$y_n|t_n, \theta, \boldsymbol{w} \propto \begin{cases} \delta(y_n > 0)\mathcal{N}\big(y_n|\boldsymbol{w}^T\boldsymbol{\theta}_n, 1\big) & \text{if } t_n = 1 \\ \delta(y_n \leq 0)\mathcal{N}\big(y_n|\boldsymbol{w}^T\boldsymbol{\theta}_n, 1\big) & \text{otherwise.} \end{cases} \tag{33}$$

$$\Rightarrow E(y_n) = \begin{cases} \boldsymbol{w}^T\boldsymbol{\theta}_n + \frac{\phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)}{1-\Phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)} & \text{if } t_n = 1 \\ \boldsymbol{w}^T\boldsymbol{\theta}_n - \frac{\phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)}{\Phi(-\boldsymbol{w}^T\boldsymbol{\theta}_n)} & \text{otherwise.} \end{cases} \tag{34}$$

We now see that $p(t_n|E(y_n)) = 1$ and equation (32) simplifies to

$$E_{\boldsymbol{y}}[\log p(\boldsymbol{t}, \boldsymbol{y}|\theta, \boldsymbol{w})] \geq \sum_{n=1}^{N} \log \mathcal{N}\big(E(y_n)|\boldsymbol{w}^T\boldsymbol{\theta}_n, 1\big) \tag{35}$$

$$= -\frac{1}{2}\sum_{n=1}^{N} \big(E(y_n) - \boldsymbol{w}^T\boldsymbol{\theta}_n\big)^2 - \frac{N}{2}\log(2\pi). \tag{36}$$

We thus see that the only difference between equations (9) and (36) is that y_n is replaced by $E(y_n)$. Hence, the E-step now involves evaluating $E(y_n)$ using equation (34), in addition to evaluating the responsibilities $\gamma(z_{dk})$ using equation (19).

M-step. As the clustering model and the prior distributions are left unchanged, the M-step also remains unchanged except for $\boldsymbol{y}$ being replaced by $E(\boldsymbol{y})$ in equations (22) and (24).